stichting
mathematisch
centrum

MC

AFDELING NUMERIEKE WISKUNDE NW 90/80 SEPTEMBER
(DEPARTMENT OF NUMERICAL MATHEMATICS)

J.G. VERWER

AN ANALYSIS OF ROSENBROCK METHODS FOR NON-LINEAR STIFF INITIAL VALUE PROBLEMS

Preprint

kruislaan 413 1098 SJ amsterdam

Printed at the Mathematical Centre, 413 Kruislaan, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

1980 Mathematics subject classification: 65L05

ACM-Computing Reviews-category: 5.17

An analysis of Rosenbrock methods for non-linear stiff initial value problems [*)]

by

J.G. Verwer

ABSTRACT

The paper presents an analysis of the Rosenbrock integration method when applied to a stiff system of the form

$$(1)\qquad \begin{aligned} \dot{x} &= f(t,x,y,\varepsilon) + \varepsilon^{-1}A(t)y, \\ \dot{y} &= g(t,x,y,\varepsilon) + \varepsilon^{-1}\mu(t)By. \end{aligned}$$

This equation possesses the following desirable model properties: (a) It permits the simultaneous occurrence of smooth and transient solution components. (b) It contains a small parameter admitting a transition to arbitrarily high stiffness. (c) The Jacobian matrix has a time-dependent eigensystem. (d) It contains non-linear terms. Provided certain assumptions have been satisfied, a characteristic of (1) is that for given initial vectors $x(0) = x_0$, $y(0) = y_0$

$$(2)\qquad \|x(t,\varepsilon)\| = O(1),\ \|y(t,\varepsilon)\| = O(\varepsilon),\quad \varepsilon \to 0,\ t \in (0,T],\ T \text{ finite.}$$

Our analysis will be directed to obtaining criteria which guarantee a similar behaviour for finite sequences of Rosenbrock approximations. By way of comparison, we also pay attention to D-stability properties of the Rosenbrock method. The property of D-stability, as introduced by van Veldhuizen, applies to the first variational equation of (1).

KEY WORDS & PHRASES: *Numerical analysis, Numerical integration, Rosenbrock methods, Non-linear stiff equations*

[*)] This report will be submitted for publication elsewhere.

1. INTRODUCTION

A well-known class of numerical integration methods for stiff systems of first order ordinary differential equations is formed by the Runge-Kutta type Rosenbrock methods. Since the original paper of ROSENBROCK [16], there has been a considerable amount of research on this type of methods. Until now the emphasis has been on the development of new methods [1,2,4,5,7,8,9, 11,17,18], on the analysis of their stability functions [3,13,14], and on the study of order conditions [14,24] (in literature the name generalized Runge-Kutta method is also used). For linear problems our theoretical knowledge of Runge-Kutta-Rosenbrock methods is rather complete nowadays. For non-linear problems however, the situation is entirely different. Though various schemes have been applied to complete test sets of non-linear problems (see e.g. [9, 10]), our knowledge of their "non-linear behaviour" is still insufficient. It is the aim of the present paper to contribute to such a knowlwdge.

We shall investigate the forerunner of all Runge-Kutta-Rosenbrock methods, that is the method originally suggested in [16]. Following ideas put forward by VAN VELDHUIZEN [21] and STETTER [20], we analyze this integration method for a typical stiff equation which possesses the following *desirable model properties:*

(a) It permits the simultaneous occurrence of smooth and transient solution components.

(b) It contains a small parameter which permits a transition to arbitrarily high stiffness.

(c) The Jacobian matrix has a time-dependent eigensystem.

(d) It contains non-linear terms.

The stiff equation is given by 2 *coupled singularly perturbed differential systems* of the form

$$\begin{aligned} \dot{x} &= f(t,x,y,\varepsilon) + \varepsilon^{-1}A(t)y, \\ \dot{y} &= g(t,x,y,\varepsilon) + \varepsilon^{-1}\mu(t)By. \end{aligned} \tag{1.1}$$

A nice feature of this *model equation* is that the vector functions f and g are allowed to be *non-linear*. We do require however that they remain bounded if

$\varepsilon \to 0$. By putting $A(t) = 0$ and $\mu(t)$ constant, we obtain the stiff system whic was used by GRIEPENTROG [6] for an investigation of a certain class of implici one-step methods. We refined his model equation to (1.1) in view of points (a) and (c) mentioned above.

Provided the right hand side functions satisfy certain assumptions, a characteristic property of (1.1) is that for given initial vectors $x(0) = x_0$, $y(0) = y_0$, the solution functions $x(t,\varepsilon)$ and $y(t,\varepsilon)$ satisfy

$$(1.2) \qquad \|x(t,\varepsilon)\| = O(1), \; \|y(t,\varepsilon)\| = O(\varepsilon), \quad \varepsilon \to 0,\; t \in (0,T],\; T \text{ finite}.$$

In our analysis of the Rosenbrock method we shall concentrate on these asymptotic relations. That is, the investigation will be directed to obtaining criteria which guarantee the existence of a constant τ^*, such that for all stepsizes $\tau \in (0,\tau^*]$ the *finite* sequences of Rosenbrock approximations satisfy similar asymptotic relations.

The boundedness of a finite approximation sequence in $\varepsilon \in (0,\varepsilon_0]$, ε_0 some constant, is in fact a property which VAN VELDHUIZEN [21] defined as D-*stability*. Among others, he has investigated this property for a 2-stage Runge-Kutta-Rosenbrock discretization of a linear model equation which describes various types of couplings between smooth and transient solution components. By way of comparison we therefore also pay attention to the model equation suggested by VAN VELDHUIZEN [21] and we present a new result on D-stability.

2. THE ROSENBROCK INTEGRATION METHOD

To begin with we define the particular class of Runge-Kutta-Rosenbrock formulas we are interested in in this paper. Let

$$(2.1) \qquad \dot{X} = F(t,X), \quad X(t_0) = X_0,$$

denote the initial value problem for a stiff system of ordinary differential equations. The following *one-step,* m-*stage integration formula* is very similar to the one originally proposed by ROSENBROCK [16]:

$$
(2.2)\qquad
\begin{aligned}
X_n^{(0)} &= X_n,\\
K_n^{(j)} &= [I-\gamma_j\tau J_n^{(j)}]^{-1}F(t_n^{(j)},X_n^{(j)}), \qquad \gamma_j > 0,\ j = 0(1)m-1,\\
X_n^{(j)} &= X_n + \tau\sum_{\ell=0}^{j-1}\lambda_{j,\ell}K_n^{(\ell)}, \qquad j = 1(1)m,\\
X_{n+1} &= X_n^{(m)}, \qquad n = 0,1,\ldots\ .
\end{aligned}
$$

X_n denotes the approximation at time $t = t_n$ and $\tau > 0$ denotes the stepsize; $t_n^{(j)} = t_n + \nu_j\tau$, where ν_j satisfies the commonly imposed condition

$$(2.3)\qquad \nu_0 = 0, \quad 0 < \nu_j < 1 \text{ if } j > 0.$$

The requirement $\nu_j > 0$, $j > 0$, is necessary for obtaining order of consistency greater than 1. I represents the unit matrix and the matrices $J_n^{(j)}$ are defined by

$$
(2.4)\qquad
\begin{aligned}
J_n^{(j)} &= J(\hat{t}_n^{(j)},\hat{X}_n^{(j)}), \qquad J(t,X) = \partial F(t,X)/\partial X,\\
\hat{t}_n^{(j)} &= \sum_{\ell=0}^{j}\alpha_{j,\ell}t_n^{(\ell)} \qquad \hat{X}_n^{(j)} = \sum_{\ell=0}^{j}\alpha_{j,\ell}X_n^{(\ell)},
\end{aligned}
$$

where the parameters $\alpha_{j,\ell}$ denote real scalars.

We wish to emphasize that the greater part of the existing Runge-Kutta-Rosenbrock schemes allow, per integration step, at most one $J(t,X)$-evaluation and one matrix factorization. For our scheme this implies

$$(2.5)\qquad \alpha_{j,0} = 1, \quad \alpha_{j,\ell} = 0 \text{ if } \ell > 0 \text{ and } \gamma_j = \gamma, \quad \gamma \text{ constant.}$$

Such a restriction is always made in order to minimize the computational overhead of one single integration step. Indeed, when we have to deal with a considerable number of components in (2.1), there is much to be said for such a choice. For the sake of analysis however, we wish to start the investigation with definition (2.4). Note that condition (2.5) implies a Jacobian evaluation at the step point (t_n,X_n). Runge-Kutta-Rosenbrock schemes evaluatting $J(t,X)$ at non-step points have been discussed in [18,23]. Schemes based on time-lagging Jacobians have been reported in [17,24,25].

In our investigation we shall need information on the performance of the Rosenbrock method when applied to certain types of linear problems

$$(2.6)\qquad \dot{X} = F(t)X, \quad X(0) = X_0, \quad t \in [0,T].$$

For problem (2.6) each stage of the Rosenbrock formula can be written as

$$(2.7)\qquad X_n^{(j)} = X_n + \tau \sum_{\ell=0}^{j-1} \lambda_{j,\ell}[I - \gamma_\ell \tau \hat{F}_n^{(\ell)}]^{-1} F_n^{(\ell)} X_n^{(\ell)}, \qquad j \geq 1,$$

where $\hat{F}_n^{(\ell)} = F(\alpha_{\ell,0}t_n^{(0)}+\ldots+\alpha_{\ell,\ell}t_n^{(\ell)})$ and $F_n^{(\ell)} = F(t_n^{(\ell)})$. Each result (2.7) thus satisfies a relation of the form

$$(2.8)\qquad X_n^{(j)} = R^{(j)}(\tau F_n^{(0)},\ldots,\tau F_n^{(j-1)};\ \tau\hat{F}_n^{(0)},\ldots,\tau\hat{F}_n^{(j-1)})X_n,$$

$R^{(j)}(\cdot;\cdot)$ being a matrix. For a constant matrix F equation (2.8) reads

$$(2.9)\qquad X_n^{(j)} = R^{(j)}(\tau F)X_n,$$

$R^{(j)}(z)$ now being a rational function of the form

$$(2.10)\qquad R^{(j)}(z) = \sum_{\ell=0}^{j} N_j^{(\ell)} z^\ell / \prod_{\ell=0}^{j} (1-\gamma_\ell z).$$

For an m-stage method, $R^{(m)}$ is called *the stability function* and determines its *absolute stability region* (see [12]). The functions $R^{(j)}$, $j < m$, may be viewed upon as *internal stability functions* (cf. [22]).

3. THE SINGULARLY PERTURBED DIFFERENTIAL SYSTEM

In this section we shall try to give a more or less precise description of the (real) initial value problem

$$(3.1)\qquad \begin{aligned} \dot{x} &= f(t,x,y,\varepsilon) + \varepsilon^{-1}A(t)y, \quad x(0) = x_0, \\ \dot{y} &= g(t,x,y,\varepsilon) + \varepsilon^{-1}\mu(t)By, \quad y(0) = y_0, \end{aligned}$$

we deal with. The problem is considered on some finite interval $[0,T]$, with initial vectors x_0 and y_0 independent of ε. We confine ourselves to problems possessing a unique bounded solution on $[0,T]$ for all $\varepsilon \in (0,\varepsilon_0]$, ε_0 some suitable constant. To that end the right hand side functions are supposed to be at least continuously differentiable in all their arguments. The vector functions f and g are allowed to be non-linear. In particular, they are supposed to be bounded in ε as $\varepsilon \to 0$. Further, $f\colon [0,T] \times \mathbb{R}^{s_1} \times \mathbb{R}^{s_2} \times (0,\varepsilon_0] \to \mathbb{R}^{s_1}$ and $g\colon [0,T] \times \mathbb{R}^{s_1} \times \mathbb{R}^{s_2} \times (0,\varepsilon_0] \to \mathbb{R}^{s_2}$, where $s_1, s_2 \geq 0$. A is a t-dependent (s_1,s_2)-matrix and μ is a scalar function which is strictly positive, i.e. $\mu(t) \geq \tilde{\mu} > 0$ for all $t \in [0,T]$. Finally, B is a constant (s_2,s_2)-matrix whose spectrum $\Lambda(B)$ lies in the negative half plane $\mathbb{C}^- = \{z \mid \mathrm{Re}(z) < 0\}$. Throughout the paper it is assumed that the above mentioned properties hold. Occasionally, the solution functions will be denoted by $x(t,\varepsilon)$ and $y(t,\varepsilon)$, respectively.

Following GRIEPENTROG [6], we shall now derive a result on the behaviour of the solution functions $x(t,\varepsilon)$, $y(t,\varepsilon)$ for decreasing ε, i.e. for *increasing stiffness*. In the remainder the symbol $\|\cdot\|$ always denotes the maximum norm. It will be clear from the context which particular maximum norm is referred to. We need the following lemma [6]:

<u>LEMMA 3.1</u>. *Let* $\alpha = \max\{\mathrm{Re}(\lambda)\colon \lambda \in \Lambda(B)\} < 0$ *and* $\tau > 0$. *Then a constant* $\hat{K}$ *exists, such that* $\|\exp(\tau B)\| \leq \hat{K}\exp(\alpha\tau/2)$. □

<u>THEOREM 3.1</u>. *For all* $t \in [0,T]$ *and* $\varepsilon \in (0,\varepsilon_0]$ *the solution functions* $x(t,\varepsilon)$ *and* $y(t,\varepsilon)$ *of the initial value problem* (3.1) *satisfy*

$$\|x(t,\varepsilon)\| \leq K_0, \quad \|\dot{x}(t,\varepsilon)\| \leq K_1[\varepsilon^{-1}\exp(\tfrac{1}{2}\alpha\tilde{\mu}\varepsilon^{-1}t) + 1],$$

$$\|y(t,\varepsilon)\| \leq \tilde{K}_0[\exp(\tfrac{1}{2}\alpha\tilde{\mu}\varepsilon^{-1}t) + \varepsilon], \tag{3.2}$$

$$\|\dot{y}(t,\varepsilon)\| \leq \tilde{K}_1[\varepsilon^{-1}\exp(\tfrac{1}{2}\alpha\tilde{\mu}\varepsilon^{-1}t) + 1],$$

K_0, $\tilde{K}_0$, K_1 *and* $\tilde{K}_1$ *being positive constants independent of t and* ε.

<u>PROOF</u>. To the second equation of (3.1) there corresponds the integral equation

$$y(t,\varepsilon) = \exp(\varepsilon^{-1}M(t)B)y_0 + y^*(t,\varepsilon), \tag{3.3}$$

where

$$M(t) = \int_0^t \mu(\tau)d\tau,$$

$$y^*(t,\varepsilon) = \int_0^t \exp(\varepsilon^{-1}[M(t) - M(\tau)]B)g(\tau,x(\tau,\varepsilon),y(\tau,\varepsilon),\varepsilon)d\tau.$$

Let $\eta > 0$ be independent of t and ε, such that for all $t \in [0,T]$ and $\varepsilon \in (0,\varepsilon_0)$, $\|g(t,x(t,\varepsilon),y(t,\varepsilon),\varepsilon)\| \le \eta$. Then, using Lemma 3.1,

$$\|y^*(t,\varepsilon)\| \le \eta\hat{K} \int_0^t \exp(\tfrac{1}{2}\alpha\varepsilon^{-1}[M(t) - M(\tau)])d\tau. \tag{3.4}$$

From the mean-value theorem it follows that

$$M(t) - M(\tau) = (t-\tau)\mu(\theta) \ge (t-\tau)\tilde{\mu}, \qquad \tau < \theta < t.$$

Substitution into (3.4) and evaluation of the resulting integral shows the existence of a constant K^* such that

$$\|y^*(t,\varepsilon)\| \le K^*\varepsilon. \tag{3.5}$$

Next, substitution into (3.3) and again using Lemma 3.2 leads us to an inequality of the form

$$\|y(t,\varepsilon)\| \le \tilde{K}_0[\exp(\tfrac{1}{2}\alpha\tilde{\mu}\varepsilon^{-1}t) + \varepsilon]. \tag{3.6}$$

The inequality for $\dot{y}(t,\varepsilon)$ is now readily obtained by substitution of $y(t,\varepsilon)$ into the differential equation.

Further, to the first equation of (3.1) there corresponds the integral equation

$$x(t,\varepsilon) = x_0 + \int_0^t f(\tau,x(\tau,\varepsilon),y(\tau,\varepsilon),\varepsilon)d\tau + \varepsilon^{-1}\int_0^t A(\tau)y(\tau,\varepsilon)d\tau. \tag{3.7}$$

By using (3.6) and (3.7) the inequality $\|x(t,\varepsilon)\| \le K_0$ is easy to obtain. Finally, direct substitution of (3.6) into the first differential equation leads to the inequality for $\dot{x}(t,\varepsilon)$. □

Because of inequalities (3.2) the solution functions $x(t,\varepsilon)$ and $y(t,\varepsilon)$ may be characterized as follows. Normally the x-solution consists of a rapidly decaying transient component and a slowly varying smooth component. The smooth component determines $x(t,\varepsilon)$ everywhere outside the transient phase. The transient behaviour of $x(t,\varepsilon)$ is completely determined by the transient of the y-solution. Further, to a large extent the magnitude of the smooth component is independent of the stiffness parameter ε. For the y-solution the situation is somewhat different. Typically, it contains a rapidly decaying transient component and, provided g does not vanish, a smooth component, viz. $y^*(t,\varepsilon)$. According to (3.6) however, $\|y^*(t,\varepsilon)\| = O(\varepsilon)$ for all $t \in (0,T]$. Consequently, the magnitude of the smooth part of $y(t,\varepsilon)$ does depend on the stiffness parameter ε. If $\varepsilon \to 0$, the smooth component vanishes for all $t \in (0,T]$. Hence, in a practical situation it will be smooth x-solution in which we are mostly interested, ε being so small that the transients can be neglected and that the smooth y-solution is of less practical interest. It shall be clear now that a suitable integration method for (3.1) should generate approximations to the smooth solutions which show a similar behaviour in ε. In particular, the method should be capable to generate such approximations with some stepsize τ being independent of ε. That is, it should be possible to *let* $\varepsilon \to 0$ *for fixed* τ. This is the requirement that we shall concentrate on (see [20] for a discussion on this type of requirement in a more general setting). Some further model aspects of (3.1), in relation to the D-stability model equation, will be treated in Section 5.

4. THE PROPERTIES OF ε-BOUNDEDNESS AND ε-ACCURACY

Suppose we apply an integration method to problem (3.1) to obtain the approximation sequences $\{x_n\}$ and $\{y_n\}$, $n = 1(1)T/\tau$, τ some given stepsize. In view of Theorem 3.1, one should then require that $\{x_n\}$ and $\{y_n\}$ satisfy

$$(4.1) \qquad x_n = O(1), \quad y_n = O(\varepsilon), \quad n = 1(1)T/\tau, \quad \varepsilon \to 0,$$

for every stepsize $\tau \in (0,\tau^*]$, where τ^* is a constant independent of ε. The property $y_n = O(\varepsilon)$ is rather unusual. Therefore, to make clear how one has to interpret the relations (4.1), we shall begin with a simple, illustrative example. Consider the scalar problem

$$\dot{y} = -\varepsilon^{-1}y + e^{-t}, \quad y(0) = y_0, \qquad \varepsilon \in (0,\varepsilon_0],$$

with exact solution

$$y(t) = \frac{\varepsilon}{1-\varepsilon} e^{-t} + (y_0 - \frac{\varepsilon}{1-\varepsilon})e^{-\varepsilon^{-1}t}.$$

Application of the 1-stage Rosenbrock scheme yields

$$y_{n+1} = \frac{1+(\gamma_0-\lambda_{10})\tau\varepsilon^{-1}}{1+\gamma_0\tau\varepsilon^{-1}} y_n + \varepsilon \frac{\lambda_{10}\tau\varepsilon^{-1}e^{-t_n}}{1+\gamma_0\tau\varepsilon^{-1}} .$$

Let us consider one single integration step, say at the initial point (t_0,y_0). Typically, if $\lambda_{10} \neq \gamma_0$, y_1 satisfies

$$|y_1| \le C_1|y_0| + C_2\varepsilon,$$

C_1 and C_2 being positive constants not depending on $\varepsilon \in (0,\varepsilon_0]$ and $\tau \in (0,\tau^*]$, $\tau^* > 0$ arbitrary. In other words, we always have $y_1 = O(1)$ as $\varepsilon \to 0$ uniformly in τ. Now set $\lambda_{10} = \gamma_0$. Then

$$y_1 = \frac{\varepsilon}{\varepsilon+\gamma_0\tau} y_0 + \varepsilon \frac{\gamma_0\tau\varepsilon^{-1}}{1+\gamma_0\tau\varepsilon^{-1}},$$

and again we may write

$$|y_1| \le C_1\varepsilon|y_0| + C_2\varepsilon.$$

The constant C_2 can again be chosen independent of τ and ε. For C_1 this no longer holds, as

$$\sup_{\varepsilon\in(0,\varepsilon_0]} (\varepsilon+\gamma_0\tau)^{-1} = (\gamma_0\tau)^{-1}.$$

Thus, for τ fixed, we can write $y_1 = O(\varepsilon)$ as $\varepsilon \to 0$; but the constant implied by this relation grows with $1/\tau$ as $\tau \to 0$. Such a situation was to be expected as the constant implied by the relation $y(\tau) = O(\varepsilon)$, $\varepsilon \to 0$, also grows as $\tau \to 0$.

As observed before, in our analysis we will consider (τ,ε)-values where $\tau \in (0,\tau^*]$ and $\varepsilon \in (0,\varepsilon_0]$, τ^* and ε_0 being constants. However, we shall always *let* $\varepsilon \to 0$ *for fixed* τ-*values*. In other words, the relations (4.1) which we are going to investigate are not valid uniformly in τ (cf. [20], p. 102). This should not be considered as an essential restriction. In the present investigation limit processes in $\tau/\varepsilon \to 0$ are of no relevance. Finally, it is worth noting that if $y_n = O(\varepsilon)$, the next approximation y_{n+1} is always $O(\varepsilon)$.

We shall now proceed with the general case. For the sake of analysis it is again convenient to *investigate one single integration step*. Because of the fact that the initial values of problem (3.1) are always constant, i.e. independent of ε, we then have to distinguish 2 types of *starting points* for such a single step. We employ the following definition:

DEFINITION 4.1. Let (t,x,y) be a given point in $[0,T] \times \mathbb{R}^{s_1} \times \mathbb{R}^{s_2}$, where x and y may be parameterized by ε. This point is called ε-bounded if both x and y are $O(1)$ as $\varepsilon \to 0$. This point is called ε-accurate if $x = O(1)$ and $y = O(\varepsilon)$ as $\varepsilon \to 0$. □

Next we shall define ε-boundedness and ε-accuracy as a property of the integration method. We wish to emphasize that in these definitions we consider one single integration step at an ε-bounded starting point. When discussing the results special attention will be paid to cases where the starting point is already ε-accurate (see Remark 4.2).

DEFINITION 4.2. The m-stage Rosenbrock method (2.2) is ε-bounded on class (3.1), or possibly on a subclass, if for all problems in this class the following is true. Suppose we are given some ε-bounded starting point (t,x,y). Then a constant τ^* exists, τ^* being independent of ε, such that for all fixed $\tau \in (0,\tau^*]$ all Rosenbrock points $(t^{(j)},x^{(j)},y^{(j)})$, $j = 1(1)m$, are again ε-bounded. □

DEFINITION 4.3. The m-stage Rosenbrock method is ε-accurate if it is ε-bounded and if, in addition, the points $(t^{(j)},x^{(j)},y^{(j)})$, $j = 1(1)m$, are ε-accurate. □

Obviously, if a method is ε-accurate relations (4.1) are satisfied. The reason why we impose our conditions also on all intermediate stages is that we deal with a *multi-stage method and a non-linear equation*. If for some j, $j < m$, the corresponding j-stage method is not ε-accurate, contrary to the m-stage method, the derivative evaluations $f(t^{(j)},x^{(j)},y^{(j)},\varepsilon)$ and $g(t^{(j)},x^{(j)},y^{(j)},\varepsilon)$ will cause the m-stage results to be unnecessarily inaccurate. Here the relevance of the non-linearity of f and g becomes apparent. For the sake of completeness we still observe that the extra conditions in Definition 4.2 might be omitted, that is, normally the ε-boundedness of $(t^{(m)},x^{(m)},y^{(m)})$ cannot be obtained if one or more intermediate points $(t^{(j)},x^{(j)},y^{(j)})$ are not ε-bounded. In view of our definition of ε-accuracy however, we prefer the extra conditions also in our definition of ε-boundedness. Furthermore, it slightly facilitates the analysis.

REMARK 4.1. The criteria of ε-boundedness and ε-accuracy are not directly related to the *propagation* of errors, such as the concept of absolute stability. They may be viewed upon as local accuracy criteria. If they are not fulfilled, we may find large local errors in non-limit situations. From this point of view ε-boundedness is similar to the criterion of D-stability proposed by VAN VELDHUIZEN [21]. They help us to distinguish between methods with the same domain of absolute stability. A more comprehensive discussion concerning the interpretation is postponed to Remark 4.2, immediately after the first result. □

Let us now investigate under what conditions a Rosenbrock method is ε-bounded and ε-accurate. First we introduce some notations. We shall write

$$X = [x,y]^T, \quad F(t,X,\varepsilon) = V(t,X,\varepsilon) + W(t,X,\varepsilon),$$

where

$$V(t,X,\varepsilon) = \begin{bmatrix} f(t,x,y,\varepsilon) \\ g(t,x,y,\varepsilon) \end{bmatrix}, \quad W(t,X,\varepsilon) = \varepsilon^{-1}\begin{bmatrix} A(t)y \\ \mu(t)By \end{bmatrix},$$

and

$$(4.2) \qquad I - \gamma\tau J(t,X,\varepsilon) = L(t,X,\varepsilon) + M(t,\varepsilon),$$

where L and M represent the block matrices

$$L(t,X,\varepsilon) = -\gamma\tau[L_{ij}(t,X,\varepsilon)]_{\substack{i=1,2\\j=1,2}}, \quad L_{11} = \frac{\partial f}{\partial x}, \quad \text{and so on,}$$

$$M(t,\varepsilon) = \begin{bmatrix} I_1 & -\gamma\tau\varepsilon^{-1}A(t) \\ 0 & I_2-\gamma\tau\varepsilon^{-1}\mu(t)B \end{bmatrix}, \quad I_i \text{ is } (s_i,s_i)\text{-unit matrix.}$$

Furthermore we (formally) denote

$$P(t,\hat{t},X,\hat{X},\varepsilon) = [L(\hat{t},\hat{X},\varepsilon)+M(\hat{t},\varepsilon)]^{-1}V(t,X,\varepsilon),$$

$$(4.3) \qquad Q(t,\hat{t},X,\hat{X},\varepsilon) = [L(\hat{t},\hat{X},\varepsilon)+M(\hat{t},\varepsilon)]^{-1}W(t,X,\varepsilon),$$

$$K(t,\hat{t},X,\hat{X},\varepsilon) = P(t,\hat{t},X,\hat{X},\varepsilon) + Q(t,\hat{t},X,\hat{X},\varepsilon).$$

Upper indices will be used in the same way as in equation (2.2). Subindices will be omitted.

Next we prove some auxiliary results on the inversion of matrices of type (4.2).

LEMMA 4.1. *A constant* $C > 0$ *exists such that* $\|M^{-1}(t,\varepsilon)\| \le C$ *for all* $t \in [0,T]$, $\varepsilon \in (0,\varepsilon_0]$ *and all* $\tau > 0$. *Moreover,* C *can be chosen independent of* t, τ *and* ε.

PROOF. Because $\gamma\tau\varepsilon^{-1}\mu(t) > 0$ and $\Lambda(B) \subset \mathbb{C}^-$, the inverse of $I_2 - \gamma\tau\varepsilon^{-1}\mu(t)B$ always exists, which implies the existence of the inverse of $M(t,\varepsilon)$. The assertion now easily follows from the definition of $M^{-1}(t,\varepsilon)$:

$$(4.4) \qquad M^{-1}(t,\varepsilon) = \begin{bmatrix} I_1 & \gamma\tau\varepsilon^{-1}A(t)[I_2-\gamma\tau\varepsilon^{-1}\mu(t)B]^{-1} \\ 0 & [I_2-\gamma\tau\varepsilon^{-1}\mu(t)B]^{-1} \end{bmatrix} \qquad \square$$

Observe that this lemma allows us to write

$$(4.2') \qquad L(t,X,\varepsilon) + M(t,\varepsilon) = M(t,\varepsilon)[I+M^{-1}(t,\varepsilon)L(t,X,\varepsilon)],$$

and, for all $t \in [0,T]$ and all $\tau > 0$,

$$(4.4') \qquad M^{-1}(t,\varepsilon) = \begin{bmatrix} I_1 & O(1) \\ 0 & O(\varepsilon) \end{bmatrix}, \quad \varepsilon \to 0.$$

Here it is once more emphasized that the $O(\varepsilon)$ relation in the above equation is not valid uniformly in τ, that is, the constant implied varies as $1/\tau$ as $\tau \to 0$. Here, and in the following, τ is kept fixed as $\varepsilon \to 0$.

LEMMA 4.2. *Let* $X = X(\varepsilon)$ *be bounded in* $\varepsilon \in (0,\varepsilon_0]$. *Then the matrix* $I+M^{-1}(t,\varepsilon)L(t,X,\varepsilon)$ *is invertible for all* $t \in [0,T]$, $\varepsilon \in (0,\varepsilon_0]$ *and* $\tau \in (0,\tau^*]$, τ^* *being a constant independent of* t *and* ε.

PROOF. Because X is bounded in $\varepsilon \in (0,\varepsilon_0]$, a constant $\tilde{C} > 0$ exists such that $\|L(t,X,\varepsilon)\| \le \tau\tilde{C}$, the constant $\tilde{C}$ being independent of $t \in [0,T]$ and $\varepsilon \in (0,\varepsilon_0]$. By making use of Lemma 4.1 we thus have $\|M^{-1}(t,\varepsilon)L(t,X,\varepsilon)\| \le \tau C\tilde{C}$, C and $\tilde{C}$ being independent of $t \in [0,T]$ and $\varepsilon \in (0,\varepsilon_0]$. The proof is completed by choosing τ^* such that $\tau^* C\tilde{C} < 1$. □

A corollary of Lemma 4.2 is that for a given $\varepsilon \in (0,\varepsilon_0]$ the *existence* of some finite approximation sequence $\{X_0,\dots,X_{T/\tau}\}$ can always be guaranteed. Further, for all $t \in [0,T]$ it is possible to write

$$(4.5) \qquad [I+M^{-1}(t,\varepsilon)L(t,X,\varepsilon)]^{-1} = \begin{bmatrix} O(1) & O(1) \\ O(\varepsilon) & I_2+O(\varepsilon) \end{bmatrix}, \quad \varepsilon \to 0,$$

provided X is bounded in ε and the stepsize τ is sufficiently small. Under these assumptions we can also write

$$(4.6) \qquad [L(t,X,\varepsilon)+M(t,\varepsilon)]^{-1} = \begin{bmatrix} O(1) & O(1) \\ O(\varepsilon) & O(\varepsilon) \end{bmatrix}, \quad \varepsilon \to 0.$$

LEMMA 4.3. *Let* $X = X(\varepsilon)$, $\hat{X} = \hat{X}(\varepsilon)$ *be bounded in* $\varepsilon \in (0,\varepsilon_0]$. *Then a constant* τ^* *exists such that for all* $\tau \in (0,\tau^*]$ *and* $t,\hat{t} \in [0,T]$, *the vector* P *satisfies*

$$P(t,\hat{t},X,\hat{X},\varepsilon) = [O(1),O(\varepsilon)]^T, \qquad \varepsilon \to 0, \tag{4.7}$$

while τ^* *does not depend on* t *and* $\hat{t}$.

PROOF. See relation (4.6) and the definition of $P(t,\hat{t},X,\hat{X},\varepsilon)$. □

Finally, our last auxiliary result:

LEMMA 4.4. *For all* $t,\hat{t} \in [0,T]$, $\varepsilon \in (0,\varepsilon_0]$ *and* $\tau > 0$ *we can write*

$$M^{-1}(\hat{t},\varepsilon)W(t,X,\varepsilon) = \varepsilon^{-1}\begin{bmatrix} A(t)y+\gamma\tau\varepsilon^{-1}A(\hat{t})[I_2-\gamma\tau\varepsilon^{-1}\mu(\hat{t})B]^{-1}\mu(t)By \\ [I_2-\gamma\tau\varepsilon^{-1}\mu(\hat{t})B]^{-1}\mu(t)By \end{bmatrix}. \tag{4.8}$$

If $A(\hat{t}) = A(t)$ *and* $\mu(\hat{t}) = \mu(t)$, *equation* (4.8) *simplifies to*

$$M^{-1}(\hat{t},\varepsilon)W(t,X,\varepsilon) = \varepsilon^{-1}\begin{bmatrix} A(t)[I_2-\gamma\tau\varepsilon^{-1}\mu(t)B]^{-1}y \\ [I_2-\gamma\tau\varepsilon^{-1}\mu(t)B]^{-1}\mu(t)By \end{bmatrix}. \tag{4.8'}$$

PROOF. See Lemma 4.1 and the definition of $W(t,X,\varepsilon)$. □

We are now ready to discuss the main results. For clarity, first a theorem on the most simple Rosenbrock scheme:

THEOREM 4.1. *The* 1-*stage Rosenbrock method is* ε-*bounded on problem class* (3.1). *This method is* ε-*accurate on this class, iff its stability function* $R^{(1)}(z)$ *satisfies* $R^{(1)}(\infty) = 0$.

PROOF. For some given ε-bounded starting point (t,X), the 1-stage scheme can be formulated as

$$\begin{aligned} X^{(1)} &= X + \tau\lambda_{10}K^{(0)}, \\ K^{(0)} &= P(t,t,X,X,\varepsilon) + Q(t,t,X,X,\varepsilon). \end{aligned} \tag{4.9}$$

According to Lemma 4.3, a constant τ^* exists such that for all $\tau \in (0,\tau^*]$, P satisfies

$$(4.10) \qquad P(t,t,X,X,\varepsilon) = [O(1),O(\varepsilon)]^T.$$

According to equation (4.8'), we can always write

$$(4.11) \qquad M^{-1}(t,\varepsilon)W(t,X,\varepsilon) = \begin{bmatrix} O(1) \\ \varepsilon^{-1}[I_2-\gamma_0\tau\varepsilon^{-1}\mu(t)B]^{-1}\mu(t)By \end{bmatrix} .$$

Application of relation (4.5) then yields, for $\tau \in (0,\tau^*]$,

$$(4.12) \qquad Q(t,t,X,X,\varepsilon) = \begin{bmatrix} O(1) \\ \varepsilon^{-1}[I_2-\gamma_0\tau\varepsilon^{-1}\mu(t)B]^{-1}\mu(t)By + O(\varepsilon) \end{bmatrix} ,$$

so that, for $\tau \in (0,\tau^*]$,

$$(4.13) \qquad X^{(1)} = \begin{bmatrix} O(1) \\ R^{(1)}(\tau\varepsilon^{-1}\mu(t)B)y + O(\varepsilon) \end{bmatrix} ,$$

where $R^{(1)}(z) = (1+(\lambda_{10}-\gamma_0)z)/(1-\gamma_0 z)$. The assertions of the theorem now follow immediately from relation (4.13). □

REMARK 4.2. In definitions (4.2) - (4.3) we consider ε-bounded starting points. Now suppose that a given starting point is already ε-accurate (and thus certainly ε-bounded). From equation (4.13) it then follows that $y^{(1)}$ is always $O(\varepsilon)$, even if $\lambda_{10} \neq \gamma_0$. Thus in order to generate ε-accurate approximation sequences *according to relations* (4.1), we might confine ourselves to putting $\lambda_{10} = \gamma_0$, i.e. $R^{(1)}(\infty) = 0$, only in the first integration step starting at the initial point (x_0,y_0). We do recommend however to require $R^{(1)}(\infty) = 0$ also in all subsequent steps, at least in the initial phase of the integration. We support this by the following observations. In our investigations we allow arbitrarily high stiffness, that is, we let $\varepsilon \to 0$. Such a limit process facilitates the analysis, and, of course, should lead to concepts being of use in practical non-limit situations (cf. [20,21]). Concerning the approximation of the transient y-component, however, the use of such a limit process may lead to a too optimistic outlook. Though $y_1 = O(\varepsilon)$ if $R^{(1)}(\infty) = 0$, the constant implied by this relation may be rather large due to the occurrence of the transient y-component (see the example in the beginning of this section; note that if $y_0 = 0$, that constant is always of modest size). For a quick damping of the y-transient, it thus is recommended to have $R^{(1)}(\infty) = 0$ in a large number of steps.

The proof of the preceding theorem also shows that in general ε-accuracy does not imply absolute stability. Suppose that we have $y_0 = 0$. Then, according to equation (4.13), $x_1 = O(1)$, $y_1 = O(\varepsilon)$ irrespective the values of λ_{10} and γ_0.

For future reference, it will appear that the remarks given here also apply to the results we shall derive for the general m-stage Rosenbrock method. □

THEOREM 4.2. *Let the* m*-stage Rosenbrock method be* ε*-bounded on class* (3.1). *The* m+1*-stage method is then* ε*-bounded on this class, iff* (i) $\hat{t}^{(m)} = t^{(m)}$ *or* (ii) *the vector* $y^{(m)}$ *is always* $O(\varepsilon)$.

PROOF. By assumption a constant τ^* exists such that for all $\tau \in (0,\tau^*]$ the vectors $X^{(j)}$, $j = 0(1)m$, are $O(1)$ as $\varepsilon \to 0$. As

$$(4.14)\qquad X^{(j)} = X^{(0)} + \tau \sum_{\ell=0}^{j-1} \lambda_{j,\ell} K^{(\ell)}, \quad j = 1,2,\ldots,$$

the increment vectors $K^{(0)},\ldots,K^{(m-1)}$ are then also $O(1)$ as $\varepsilon \to 0$. Consequently, the m+1-stage method is ε-bounded, i.e. $X^{(m+1)} = O(1)$, iff $K^{(m)} = O(1)$. According to (4.3), $K^{(m)}$ is given by

$$(4.15)\qquad K^{(m)} = P(t^{(m)},\hat{t}^{(m)},X^{(m)},\hat{X}^{(m)},\varepsilon) + Q(t^{(m)},\hat{t}^{(m)},X^{(m)},\hat{X}^{(m)},\varepsilon).$$

We first establish, by means of Lemma 4.3, that if τ^* is sufficiently small $P^{(m)}$ always satisfies

$$(4.16)\qquad P^{(m)} = [O(1),\ O(\varepsilon)]^T.$$

We therefore can concentrate on $Q^{(m)}$ which is defined by

$$(4.17)\qquad Q^{(m)} = [I + M^{-1}(\hat{t}^{(m)},\varepsilon)L(\hat{t}^{(m)},\hat{X}^{(m)},\varepsilon)]^{-1}M^{-1}(\hat{t}^{(m)},\varepsilon)W(t^{(m)},X^{(m)},\varepsilon),$$

provided τ^* is sufficiently small (see Lemma 4.2). Let us consider the expression for $M^{-1}(\hat{t}^{(m)},\varepsilon)W(t^{(m)},X^{(m)},\varepsilon)$ as given in equation (4.8). The second component vector is certainly $O(1)$. On class (3.1) the first component vector can be seen to be $O(1)$, iff $\hat{t}^{(m)} = t^{(m)}$ or $y^{(m)} = O(\varepsilon)$. By making use of relation (4.5) we thus conclude that, on class (3.1), the first component vector of $Q^{(m)} = O(1)$, iff $\hat{t}^{(m)} = t^{(m)}$ or $y^{(m)} = O(\varepsilon)$. □

Evidently, ε-boundedness is determined by the boundedness of the expression

$$\varepsilon^{-1}A(t)y + \gamma\tau\varepsilon^{-2}A(\hat{t})[I_2 - \gamma\tau\varepsilon^{-1}\mu(\hat{t})B]^{-1}\mu(t)By. \tag{4.18}$$

Bearing this in mind the following theorem is now easy to prove:

THEOREM 4.3. *Any Rosenbrock method is ε-bounded on the class of problems* (3.1) *where* A *and* μ *are both constant. Any Rosenbrock method is ε-bounded on the class where* A(t) *is the zero matrix.* □

The requirement $\hat{t}^{(m)} = t^{(m)}$, occurring in Theorem 4.2, is rather unpleasant because it means a Jacobian evaluation at stage m (see (2.4)). We shall proceed with the second requirement. This requirement is certainly fulfilled if the m-stage method is also ε-accurate. A more detailed result is given below:

THEOREM 4.4. *Let the* m*-stage Rosenbrock method be ε-bounded on class* (3.1), *or possibly on a subclass. Then the vector* $y^{(m)}$ *is always* $O(\varepsilon)$, *iff for all functions* μ *in this class the matrix*

$$R^{(m)}(\tau\varepsilon^{-1}\mu(t^{(0)})B,\ldots;\ldots,\tau\varepsilon^{-1}\mu(\hat{t}^{(m-1)})B) = O(\varepsilon), \quad \varepsilon \to 0, \tag{4.19}$$

where $R^{(m)}$ *is defined as in equation* (2.8).

PROOF. By assumption a constant τ^* exists such that for all $\tau \in (0,\tau^*]$, the vectors $X^{(j)}$, $j = O(1)m$, are $O(1)$ as $\varepsilon \to 0$. By repeating part of the derivations in the proof of Theorem 4.2 we then can prove that, for $\tau \in (0,\tau^*]$, the second component of the vector $K^{(j)}$, say $k^{(j)}$, satisfies

$$k^{(j)} = \varepsilon^{-1}[I_2 - \gamma_j\tau\varepsilon^{-1}\mu(\hat{t}^{(j)})B]^{-1}\mu(t^{(j)})By^{(j)} + O(\varepsilon), \quad j = O(1)m-1. \tag{4.20}$$

By definition

$$\begin{aligned} y^{(m)} &= y + \tau \sum_{j=0}^{m-1} \lambda_{m,j}\, k^{(j)} \\ &= y + \tau \sum_{j=0}^{m-1} \lambda_{m,j}\varepsilon^{-1}[I_2-\gamma_j\tau\varepsilon^{-1}\mu(\hat{t}^{(j)})B]\mu(t^{(j)})By^{(j)} + O(\varepsilon). \end{aligned} \tag{4.21}$$

According to relations (2.7) - (2.8) we thus have

$$y^{(m)} = R^{(m)}(\tau\varepsilon^{-1}\mu(t^{(0)})B,\dots\,;\,\dots,\tau\varepsilon^{-1}\mu(\hat{t}^{(m-1)})B)y + O(\varepsilon). \quad \square \tag{4.22}$$

By making use of Theorems 4.1 - 4.2, and by repeated application of Theorem 4.4, we next can prove

THEOREM 4.5. *The* m-*stage Rosenbrock method is* ε-*accurate on class* (3.1), *or possibly on a subclass, iff for all functions* μ *in this class the matrices* $R^{(j)}$, j = 1(1)m, *satisfy*

$$R^{(j)}(\tau\varepsilon^{-1}\mu(t^{(0)})B,\dots\,;\,\dots,\tau\varepsilon^{-1}\mu(\hat{t}^{(j-1)})B) = O(\varepsilon), \qquad \varepsilon \to 0. \quad \square \tag{4.23}$$

It thus depends on all matrices $R^{(j)}$, j = 1(1)m, whether an m-stage method is ε-accurate or not. We shall show that conditions (4.23) can be satisfied, if and only if all stability functions $R^{(j)}(z)$ do have a zero at infinity.

THEOREM 4.6. *The* m-*stage Rosenbrock method is* ε-*accurate on class* (3.1), *if and only if the stability function* $R^{(m)}(z)$, *as well as all internal stability functions* $R^{(j)}(z)$, j = 1(1)m-1, *do have a zero at infinity.*

PROOF. The matrices $R^{(j)}$ appearing in equation (4.23) are defined by the recursion

$$\begin{aligned} R^{(0)} &= I_2, \\ R^{(j)} &= I_2 + \sum_{\ell=0}^{j-1} \lambda_{j,\ell}[I_2 - \gamma_\ell\tau\varepsilon^{-1}\mu(\hat{t}^{(\ell)})B]^{-1}\tau\varepsilon^{-1}\mu(t^{(\ell)})BR^{(\ell)}. \end{aligned} \tag{4.24}$$

Because $\hat{t}^{(0)} = t^{(0)}$, $R^{(j)}$, for all $j > 0$, can be rewritten as

$$(4.25) \qquad R^{(j)} = \frac{I_2+(\lambda_{j,0}-\gamma_0)\tau\varepsilon^{-1}\mu(t^{(0)})B}{I_2-\gamma_0\tau\varepsilon^{-1}\mu(t^{(0)})B} + \sum_{\ell=1}^{j-1} \frac{\lambda_{j,\ell}\tau\varepsilon^{-1}\mu(t^{(\ell)})B}{I_2-\gamma_\ell\tau\varepsilon^{-1}\mu(\hat{t}^{(\ell)})B} R^{(\ell)} .$$

Now suppose that $\lambda_{j,0} = \gamma_0$ for $j = 1(1)m$. It is then immediate that all matrices $R^{(j)}$, $j = 1(1)m$, are $O(\varepsilon)$. Next suppose that for at least one j, $\lambda_{j0} \neq \gamma_0$. Let j_1 be the minimal integer from these. Then $R^{(j)} = O(\varepsilon)$, $j = 1(1)j_1-1$ and thus $R^{(j_1)}$ must be $O(1)$. Consequently, all matrices $R^{(j)}$ are $O(\varepsilon)$ if and only if all parameters $\lambda_{j,0}$ satisfy $\lambda_{j,0} = \gamma_0$. Finally, by again making use of relation (4.24), it follows that all rational functions $R^{(j)}(z)$ satisfy $R^{(j)}(\infty) = 0$, if and only if $\lambda_{j,0} = \gamma_0$, $j \leq m$. □

We conclude this section by giving a *short summary of our main results*. First we establish that any Rosenbrock method (2.2) is ε-bounded on the 2 classes of problems where $A(t) = 0$ and A and μ are both constant, respectively. An m-stage method using m-evaluations of the Jacobian matrix within one single integration step, i.e. $\hat{t}^{(j)} = t^{(j)}$ for $j = 0(1)m-1$, is always ε-bounded. Apart from the fact that we prefer ε-accuracy, such a method will usually be computationally expensive. To guarantee that an m-stage Rosenbrock method is ε-accurate, and thus ε-bounded, we have to require that the stability function $R^{(m)}(z)$, as well as all internal stability functions $R^{(j)}(z)$, do have a zero at infinity. For schemes using at most 1 Jacobian evaluation per step such internal stability functions are in general also necessary to obtain ε-boundedness, i.e., if $R^{(m)}(\infty) = 0$, ε-boundedness generally implies ε-accuracy. These facts confirm the relevance of the concept of internal stability introduced in [22]. Finally it is once more noted that in Definitions 4.2 - 4.3 we consider ε-bounded starting points (t,x,y). If we perform an integration step at an ε-accurate point, all Rosenbrock points $(t^{(j)},x^{(j)},y^{(j)})$ are always ε-accurate again. This can be shown by means of relation (4.22) which can be written down for $j = 1,2,\ldots$. See Remark 4.2 for a correct interpretation of ε-accurate starting points.

5. SOME NOTES ON D-STABILITY

The boundedness of a finite approximation sequence in $\varepsilon \in (0,\varepsilon_0]$ is a property which, in a somewhat different setting, has earlier been defined as D-stability (see VAN VELDHUIZEN [21]). Among others, van Veldhuizen has investigated Rosenbrock methods (a 2-stage generalized Runge-Kutta method). In this section we shall connect our concepts of ε-boundedness and ε-accuracy with van Veldhuizen's D-stability concept. Further we shall point out a certain shortcoming of our model equation. The section is concluded with a new result on D-stability.

D-stability applies to linear equations of type (2.6). Let S represent some class of stiff systems of type (2.6). For our Rosenbrock method (2.2) the definition of D-stability, related to the class S, is then given by

DEFINITION 5.1. The m-stage Rosenbrock method is called $D(S)$-stable if for all stiff systems in the class S, for all $t \in [0,T]$, and all $\tau \in (0,\tau^*]$, the matrix $R^{(m)}$ given in equation (2.8) satisfies

$$(5.1) \qquad \| R^{(m)}(\tau F^{(0)},\ldots,\tau F^{(m-1)};\ \tau\hat{F}^{(0)},\ldots,\tau\hat{F}^{(m-1)}) \| \le M < \infty,$$

M a constant depending only on τ^* and the class S. □

For S_1 being the class of all linear, homogeneous equations contained in class (3.1), the definition of $D(S_1)$-stability slightly differs from our definition of ε-boundedness. In Definition (5.1) no additional conditions on the intermediate stages occur. Further, as D-stability applies to linear problems, it does not take account of the ε-dependence of starting points. Of course, much depends on the choice of S. In [21] the investigation is concentrated on

DEFINITION 5.2. The class S consists of all linear systems $\dot{X} = F(t)X$, parameterized by $\varepsilon \in (0,\varepsilon_0]$, that satisfy the conditions

(S1) $X(t) \in \mathbb{C}^2$,

(S2) $F(t) = E(t)D(t)E^{-1}(t)$, for all $t \in [0,T]$, where

$$D(t) = \begin{bmatrix} d_1(t) & 0 \\ 0 & \varepsilon^{-1} d_2(t) \end{bmatrix}, \quad \mathrm{Re}(d_2(t)) \le \tilde{d}_2 < 0 \text{ for all } t \in [0,T].$$

(S3) d_1, d_2, E and E^{-1} depend smoothly on t and possibly on $\varepsilon \in (0,\varepsilon_0]$, and the derivatives from order zero up to a sufficiently high order are bounded on $[0,T] \times (0,\varepsilon_0]$. □

Any system belonging to S thus is of the form

$$(5.2) \qquad \dot{x} = \varepsilon^{-1} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} x,$$

where a_{ij} depends smoothly on t and ε (see [21] for a discussion of properties (S1) - (S3)). Now consider our model equation (3.1). Let, for reasons of comparison, both x and y be scalars and complex (for our investigation a non-essential change). Next, linearization of f and g yields systems of the form

$$(5.3) \qquad \dot{x} = \begin{bmatrix} a_{11} & \varepsilon^{-1} a_{12} \\ a_{21} & \varepsilon^{-1} a_{22} \end{bmatrix} x,$$

where a_{ij} again depends smoothly on t and, possibly, on ε. We shall further assume that all systems (5.3) satisfy properties (S1) - (S3). For convenience we introduce the class S_2:

DEFINITION 5.3. The class S_2 consists of all linear systems (5.3) that satisfy properties (S1) - (S3). □

Equation (5.3) may be viewed upon as a prototype of the first variational form of our non-linear model equation.

Obviously, S_2 is a subclass of class S. In fact, when considered as a model equation, equation (5.3), and also equation (3.1), exhibits a certain shortcoming. We shall explain this. According to point (a) mentioned in the introduction, a desirable property of a model equation is that it permits the simultaneous occurrence of smooth and transient solution components. Furthermore, given such a model equation, it is desirable that these components may be coupled in various ways. Following van Veldhuizen, we shall shortly describe these couplings. Consider a system belonging to class $\mathcal{S}$. Denote $Y(t) = E^{-1}(t)X(t)$. Then

$$(5.4) \qquad \dot{Y} = [D(t) - C(t)]Y, \quad C(t) = E^{-1}(t)\dot{E}(t),$$

is equivalent to this system. In case $C(t)$ is a diagonal matrix for all $t \in [0,T]$, it has been uncoupled by the transformation $X = EY$, that is, there exists no coupling between smooth and transient solution components. In case a coupling exists, we employ the following definition:

DEFINITION 5.4. The coupling from the smooth to the transient component, at $t = t^*$, is weak if $C_{21}(t^*) = O(\varepsilon)$. The coupling from the transient to the smooth component, at $t = t^*$, is weak if $C_{12}(t^*) = O(\varepsilon)$. If a coupling is not weak, we call it strong. □

It is perhaps clarifying to remark that, as a consequence of property (S3), $C(t)$ remains bounded on $[0,T]$ as $\varepsilon \to 0$. Now given a problem from class $\mathcal{S}_2 \subset \mathcal{S}$, the following lemma reveals the couplings it describes:

LEMMA 5.1. *Let* $\dot{X} = F(t)X$ *belong to class* $\mathcal{S}_2$.

(i) *Suppose* $F(t)$ *is non-triangular. Then*

$$(5.5) \qquad \begin{aligned} C_{12}(t) &= \frac{d_2(t)\dot{a}_{12}(t)-\dot{d}_2(t)a_{12}(t)-\varepsilon\,\frac{d}{dt}\,(a_{11}(t)a_{12}(t))}{[d_2(t)-\varepsilon d_1(t)]a_{12}(t)}, \\ C_{21}(t) &= \frac{\varepsilon[\dot{d}_1(t)a_{12}(t)-d_1(t)\dot{a}_{12}(t)]-\varepsilon\,\frac{d}{dt}\,(a_{11}(t)a_{12}(t))}{[d_2(t)-\varepsilon d_1(t)]a_{12}(t)} = O(\varepsilon). \end{aligned}$$

(ii) *Suppose* $F(t)$ *is lower triangular. Then*

$$(5.6) \qquad \begin{aligned} C_{12}(t) &= 0, \\ C_{21}(t) &= \frac{\varepsilon a_{21}(t)[\dot{a}_{22}(t)-\varepsilon\dot{a}_{11}(t)]+\varepsilon\dot{a}_{21}(t)[\varepsilon a_{11}(t)-a_{22}(t)]}{\varepsilon a_{11}(t)-a_{22}(t)} = O(\varepsilon). \end{aligned}$$

(iii) *Suppose* F(t) *is upper triangular. Then*

(5.7)
$$c_{12}(t) = \frac{a_{22}(t)\dot{a}_{12}(t)-a_{12}(t)\dot{a}_{22}(t)+\varepsilon a_{12}(t)[\dot{a}_{11}(t)-\dot{a}_{22}(t)]}{a_{22}(t)-\varepsilon a_{11}(t)},$$

$$c_{21}(t) = 0.$$

PROOF. By making use of properties (S1) - (S3) the proof follows easily from elementary matrix algebra. □

We can conclude that in all cases $c_{21}(t) = O(\varepsilon)$. Consequently, *class* S_2 *does not describe a strong coupling from smooth to transient components.* This is in fact the shortcoming we meant. Furthermore, class S_2 does not contain all systems having $c_{21}(t) = O(\varepsilon)$, so $S_2 \subset W_{st}$, W_{st} being the subclass of S for which on the whole time interval $c_{21}(t) = O(\varepsilon)$ (cf. [21]).

The importance of the coupling classification is clearly exemplified by Theorem 3.1 in [21]. This theorem applies to our 2-stage Rosenbrock method satisfying restriction (2.5). Let W_{ts} be defined in the same way as W_{st}. One of the results of van Veldhuizen's theorem then is that this 2-stage method is $D(W_{ts} \cup W_{st})$-stable, iff $R^{(1)}(\infty) = 0$ (cf. our Theorems 4.1 - 4.2). A negative result is that this 2-stage method, and probably any multistage method satisfying restriction (2.5) , cannot be $D(S)$-stable. So, if we have to deal with a problem possessing a strong coupling from smooth to transient, as well as from transient to smooth, the property of D-stability cannot be guaranteed For the sake of clarity we shall give a simple example:

EXAMPLE 5.1. Consider the system (a similar example has been quoted by KREISS [11])

(5.8)
$$\dot{x} = E(t)\begin{bmatrix} d_1(t) & 0 \\ 0 & \varepsilon^{-1}d_2(t) \end{bmatrix} E^{-1}(t),$$

where d_1 and d_2 satisfy property (S3) and where

(5.9)
$$E(t) = \begin{bmatrix} \cos \nu t & -\sin \nu t \\ \sin \nu t & \cos \nu t \end{bmatrix}, \ \nu \text{ constant.}$$

For the new dependent variable $Y(t) = E^{-1}(t)X(t)$ we obtain (cf. equation (5.4))

$$(5.10) \qquad \dot{Y} = \begin{bmatrix} d_1(t) & \nu \\ -\nu & \varepsilon^{-1}d_2(t) \end{bmatrix} Y.$$

Clearly, for all t, $C_{12}(t) = -\nu$ and $C_{21}(t) = \nu$. Thus we have to deal with strong couplings, i.e. equation (5.8) belongs to class $\mathcal{S}\backslash\{\mathcal{W}_{st} \cup \mathcal{W}_{ts}\}$. Furthermore, for our 2-stage Rosenbrock method satisfying restriction (2.5), it is not difficult to verify that the constant M in Definition 5.1 does not exist, i.e. the property of D-stability does not hold. It is also worthwhile to observe that system (5.10) is of type (5.3), that is, it belongs to class $\mathcal{S}_2$. Application of Theorem 4.3 then immediately shows (A(t) is zero) that any Rosenbrock method generates finite ε-bounded approximation sequences for system (5.10). We must conclude that a simple transformation of the differential equaion may lead to a qualitatively different behaviour of the Rosenbrock method. □

D($\mathcal{S}$)-stability can be proved if we drop restriction (2.5):

<u>THEOREM 5.1</u>. *The* m-*stage Rosenbrock method* (2.2) *is* D($\mathcal{S}$)-*stable if* $\hat{t}^{(j)} = t^{(j)}$, *i.e.* $\hat{X}^{(j)} = X^{(j)}$, *for* $j = 0(1)m-1$.

<u>PROOF</u>. For the linear equation $\dot{X} = F(t)X$, the expression for $X^{(j)}$ now reads (see equation (2.7))

$$(5.11) \qquad X^{(j)} = X + \tau \sum_{\ell=0}^{j-1} \lambda_{j,\ell}[I-\gamma_\ell \tau F^{(\ell)}]^{-1}F^{(\ell)}X^{(\ell)}, \qquad j = 1(1)m.$$

If $\dot{X} = F(t)X$ belongs to class $\mathcal{S}$, then

$$(5.12) \qquad [I-\gamma_\ell \tau F^{(\ell)}]^{-1}\tau F^{(\ell)} = E^{(\ell)}[I-\gamma_\ell \tau D^{(\ell)}]^{-1}\tau D^{(\ell)}[E^{(\ell)}]^{-1}.$$

Hence a constant $\tau^{(\ell)} > 0$ exists such that, uniform in $\tau \in (0,\tau^{(\ell)}]$, the matrix (5.12) is O(1) as $\varepsilon \to 0$. □

By definition, $\hat{t}^{(0)} = t^{(0)}$. Thus the 1-stage scheme is always D(S)-stable. For a multistage scheme the equalities $\hat{t}^{(j)} = t^{(j)}$ imply m Jacobian matrix evaluations per integration step. It shall be clear that in most practical situations the use of such a D(S)-stable Rosenbrock scheme involves a considerable computational overhead.

As pointed out by VAN VELDHUIZEN [21], the lack of D-stability may result in large local errors in non-limit situations. Therefore, the conclusion that can be drawn from this section is that Rosenbrock schemes are not the proper methods for problems exhibiting only strong couplings. In case of a weak coupling, we expect that D-stability can always be obtained if all *internal* stability functions $R^{(j)}(z)$ satisfy $R^{(j)}(\infty) = 0$ (cf. our Theorem 4.6).

6. FINAL CONCLUSIONS

The purpose of the present investigation was to get insight in the performance of Runge-Kutta-Rosenbrock methods when applied to *non-linear* stiff systems. To that end we analyzed the original Rosenbrock method for 2 model systems. Among others, these models have in common a time-varying eigensystem, like a general non-linear equation has. For schemes using one Jacobiar matrix evaluation per time step our results strongly indicate to the conclusion that to cope with time-varying eigensystems, in any case in the initial phase of the integration, the stability function $R^{(m)}(z)$, as well as all internal stability functions $R^{(j)}(z)$, should have a zero at infinity (see also [22]). Furthermore, if we have to deal with eigensystems exhibiting only strong couplings, the performance of these Rosenbrock schemes will be far from optimal. For such problems one should reevaluate the Jacobian at each stage. Unfortunately, this strategy will normally result in a considerable computational overhead.

As observed before, the properties we investigated are not directly related to the propagation of errors, like, for example, A-stability does. Hence if a scheme is constructed on the basis of our results, the resulting stability function $R^{(m)}(z)$ needs an additional investigation.

In a future contribution the author intends to report numerical experiments with the aim of clarifying and supporting the results and conclusions of the present investigation.

ACKNOWLEDGEMENT

The author wishes to acknowledge M. van Veldhuizen for his valuable comments on an earlier version of the paper.

REFERENCES

[1] BUI, T.D., *Some A-stable and L-stable methods for the numerical integration of stiff ordinary differential equations,* JACM, 26 (1979) pp. 483-493.

[2] BUI, T.D. & T.R. BUI, *Numerical methods for extremely stiff systems of ordinary differential equations,* Appl. Math. Modelling, 3 (1979) pp. 335-358.

[3] BURRAGE, K., *A special family of Runge-Kutta methods for solving stiff differential equations,* BIT, 18 (1978) pp. 22-41.

[4] CASH, J.R., *Semi-implicit Runge-Kutta procedures with error estimates for the numerical integration of stiff systems of ordinary differential equations,* JACM, 23 (1976) pp. 455-460.

[5] FRIEDLI, A., *Verallgemeinerte Runge-Kutta-Verfahren zur Lösung steifer Differentialgleichungssysteme,* Lecture Notes in Mathematics 631, Springer-Verlag, Berlin, pp. 35 - 50, 1977.

[6] GRIEPENTROG, E., *Numerische Integration steifer Differentialgleichungssysteme mit Einschrittverfahren,* Beiträge zur Numerische Mathematik, 8 (1980) pp. 59-74.

[7] HOUWEN, P.J. VAN DER, *Construction of integration formulas for initial value problems,* North-Holland Publishing Company, Amsterdam, 1977.

[8] KAPS, P., *Modifizierte Rosenbrockmethoden der Ordnung 4, 5 und 6 zur numerischen Integration steifer Differentialgleichungen,* Dissertation Universität Innsbruck, 1977.

[9] KAPS, P. & P. RENTROP, *Generalized Runge-Kutta methods of order 4 with stepsize control for stiff ordinary differential equations,* Numer. Math., 33 (1979) pp. 55-68.

[10] KAPS, P. & G. WANNER, *A study of Rosenbrock-type methods of high order,* Report Universität Innsbruck, 1979.

[11] KREISS, H.O., *Difference methods for stiff ordinary differential equations,* SIAM J. Numer. Anal., 15 (1978) pp. 21-58.

[12] LAMBERT, J.D., *Computational methods in ordinary differential equations,* John Wiley & Sons, London, 1973.

[13] NØRSETT, S.P. & G. WANNER, *The real-pole sandwich for rational approximations and oscillation equations,* BIT, 19 (1979) pp. 79-94.

[14] NØRSETT, S.P. & A. WOLFBRANDT, *Attainable order of rational approximations to the exponential function with only real poles,* BIT, 17 (1977) pp. 200-208.

[15] NØRSETT, S.P. & A. WOLFBRANDT, *Order conditions for Rosenbrock type methods,* Numer. Math., 32 (1979) pp. 1-15.

[16] ROSENBROCK, H.H., *Some general implicit processes for the numerical solution of differential equations,* Computer J., 18 (1964) pp. 50-64.

[17] SCHOLZ, S., *Modifizierte Rosenbrock-Verfahren mit genäherter Jacobi-Matrix,* Sektion Mathematik, Technische Universität Dresden, 1979.

[18] SCHOLZ, S., S-*stabile modifizierte Rosenbrock-Verfahren* 3 *und* 4 *Ordung,* Sektion Mathematik, Technische Universität Dresden, 1978.

[19] STEIHAUG, T. & A. WOLFBRANDT, *An attempt to avoid exact Jacobian and nonlinear equations in the numerical solution of stiff differential equations,* Math. Comp., 33 (1979) pp. 521-534.

[20] STETTER, H.J., *Towards a theory for discretizations of stiff differential systems,* Lecture Notes in Mathematics 506, Springer-Verlag, Berlin, pp. 190-201, 1976.

[21] VELDHUIZEN, M. VAN, D-*stability,* SIAM J. on Numer. Anal., to appear.

[22] VERWER, J.G., S-*stability properties for generalized Runge-Kutta methods*, Num. Math., 27 (1977) pp. 359-370.

[23] VERWER, J.G., *On generalized Runge-Kutta methods using an exact Jacobian at a non-step point,* ZAMM, 60 (1980) pp. 263-265.

[24] VERWER, J.G. & S. SCHOLZ, *Rosenbrock methods and time-lagged Jacobian matrices,* Report NW 82/80, Mathematisch Centrum, Amsterdam, 1980 (prepublication).

[25] WOLFBRANDT, A., *A study of Rosenbrock processes with respect to order conditions and stiff stability,* Thesis Chalmers Univ. of Technology, Göteborg, Sweden, 1977.